

---

# GATE Mimir and cloud services

---

Multi-paradigm indexing and search tool

Pay-as-you-go large-scale annotation

---



# GATE Mimir

---

- GATE Mimir is an indexing system for GATE documents.
- Mimir can index:
  - Text: the original document content is indexed (based on Token annotations)
  - Annotations: annotations and features
  - Semantics: annotations can be linked to external ontologies which can be used at search time
- Mimir queries allow for any combination of these



# Why Mimir?

---

- Standard search covers the text.
- GATE documents also have annotations, which give access to the document's:
  - Structure (sections, titles, etc.)
  - Linguistic features (nouns, verbs, etc.)
  - Semantics
  - Etc.
- Examples: <http://demos.gate.ac.uk/mimir>
  - BBC News demo



# Mimir query language

---

- Simplest queries are free text
  - *text*
  - “quoted string”
- Searching against the document text

# Plain text queries

Harriet Harman

Search

Documents 1 to 20 of 81:

## [UK childcare needs to be more affordable - CentreForum](#) (cached)

quality' MP **Harriet Harman** was the architect

## [Birth weight among social mobility checks - Nick Clegg](#) (cached)

's deputy leader **Harriet Harman** said Mr Clegg

## [Ed Miliband's shadow cabinet and ministerial teams](#) (cached)

Miliband Opposition leader **Harriet Harman** Deputy Leader & in 2011. **HARRIET HARMAN** - DEPUTY LEADER

## [PM's response to Skinner Commons question 'shameful'](#) (cached)

. Deputy leader **Harriet Harman** wrote on Twitter

## [Daily Politics and Sunday Politics highlights of 2012](#) (cached)

Sunday April 29 **Harriet Harman** on Hunt, Cameron and Clegg **Harriet Harman** struggles with bank on health by **Harriet Harman**  
PMQs: Harriet ...

## [Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance'](#) (cached)

Shadow Culture Secretary **Harriet Harman** says Jeremy Hunt culture secretary, **Harriet Harman**, told the

## [No Rupert Murdoch deal, says Alastair Campbell](#) (cached)



# Plain text queries

Harriet Harman says

Search

Documents 1 to 20 of 29:

**Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance'** (cached)

Shadow Culture Secretary **Harriet Harman says** Jeremy Hunt was

**Ed Miliband defends Iraq war condemnation** (cached)

, says Harman **Harriet Harman says** Labour will be

**Ed Miliband tells Labour: We're the optimists now** (cached)

, says Harman **Harriet Harman says** Labour will be

**Labour must have credible deficit plan, says Darling** (cached)

, says Harman **Harriet Harman says** Labour will be

**David Miliband says he won't join brother Ed's team** (cached)

, says Harman **Harriet Harman says** Labour will be

**Balls: Labour must fight cuts 'every inch of the way'** (cached)

, says Harman **Harriet Harman says** Labour will be

**Rob Ainsworth clashes with ex-aide over Trident** (cached)



# Token features

---

- Free text queries are actually searching the `string` features of the GATE Token annotations
- Can also search other token features, e.g. `root` (morphology) or `category` (POS)



# Morphology

Harriet Harman root:say

Search

Documents 21 to 18 of 38:

[David Cameron criticised for 'calm down dear' jibe \(cached\)](#)

former equality minister **Harriet Harman said** Mr Cameron's

[Queen's Speech: Biggest change to voter registration \(cached\)](#)

, Labour's **Harriet Harman said** the government was

[Harriet Harman struggles with bank bonus and job figures \(cached\)](#)

in Coventry, **Harriet Harman said**: "I

[PMQs: Harriet Harman and Nick Clegg on unemployment \(cached\)](#)

Labour, but **Harriet Harman said** unemployment was falling

[Leveson Inquiry: Jeremy Hunt fair on BSkyB, says top civil servant \(cached\)](#)

Shadow culture secretary **Harriet Harman said**: "David

[Jeremy Hunt: I followed due process over BSkyB \(cached\)](#)

But Labour's **Harriet Harman said** Mr Hunt had

[Ed Miliband 'will marry' but politics 'got in the way' \(cached\)](#)

, says Harman **Harriet Harman says** Labour will be



# Gaps

---

- Default combinator is *sequence* – terms must be adjacent
  - Different from typical search engines
- Can allow gaps with  $[n..m]$
- Arbitrary gap with AND
  - $x \text{ AND } y$  finds the shortest span that covers both



# Gaps

Harriet Harman [0..5] root:say

Search

Documents 1 to 20 of 43:

[Birth weight among social mobility checks - Nick Clegg](#) (cached)

's deputy leader **Harriet Harman said** Mr Clegg would

[Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance'](#) (cached)

Shadow Culture Secretary **Harriet Harman says** Jeremy Hunt was

[Jeremy Hunt quit call after Leveson BSkyB evidence](#) (cached)

Labour deputy leader **Harriet Harman also said** Mr Hunt should

[000375\\_http://www.bbc.co.uk/news/uk-14481315](http://www.bbc.co.uk/news/uk-14481315) (cached)

Miliband and deputy **Harriet Harman are in effect saying** Prime Minister David



# Annotations

---

- So far nothing a standard search engine couldn't do...
- But Mimir also indexes annotations
- Syntax:
  - `{AnnotationType feat1=val1 feat2=val2}`
  - Feature comparisons can be =, <, <=, >=, >
  - UI offers pop-up with available types/features
- Let's generalise – *any* person, not just Harriet

# Annotations

```
{Person} [0..5] root:say
```

Documents 1 to 20 of 5495:

## [Apple's Sir Jonathan Ive reaffirms desire to stay at company \(cached\)](#)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

## [Diamond Jubilee Tube train was faulty \(cached\)](#)

be happening' **Ms Siggs said**: "It

## [Warning over deep-ocean stowaways \(cached\)](#)

using the famous **Alvin sub say** the vehicle picked embarrassment. But **Dr Voight says** the experience is it," **Dr Voight said**. "We

## [School building system not fit for purpose, review says \(cached\)](#)

government-commissioned review by **Sebastian James of Dixons Group said** value for money by Education Secretary **Michael Gove, Mr James said**. Schools with shadow education secretary **Andy Burnham said** Mr Gove had ...

# Other operators

- Containment (use parentheses to group)
  - Query1 IN Query2
  - Query1 OVER Query2

```
( {Person} [0..5] root:say ) IN {Content}
```

Documents 1 to 20 of 4924:

[Apple's Sir Jonathan Ive reaffirms desire to stay at company \(cached\)](#)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

[Diamond Jubilee Tube train was faulty \(cached\)](#)

be happening' **Ms Siggs said**: "It

# Other operators

- Alternative
  - Query1 OR Query2
  - Query1 | Query2

```
{Person} | {Organization} [0..2] root:say
```

Documents 1 to 20 of 8611:

[Apple's Sir Jonathan Ive reaffirms desire to stay at company](#) (cached)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

[Diamond Jubilee Tube train was faulty](#) (cached)

for London (TfL) **has said**. Passengers were be happening' **Ms Siggs said**: "It broke down. **London Ambulance Service said** two ambulance crews ...



# Other operators

---

- Set difference
  - `Query1 MINUS Query2`
  - Returns all spans that match Query1 but are not also matches of Query2
  - E.g. sentences that *don't* mention a location

```
{Sentence} MINUS  
( {Sentence} OVER {Location} )
```



# Try it!

- <http://demos.gate.ac.uk/mimir>
  - BBC News demo
- Find:
  - Document titles
  - Date expressions
  - Amounts of money **being paid**
  - Amounts of money **being received**
- Hint: if you get too much noise, try
  - restricting to matches within a sentence
  - using `IN {Content}` to ignore boilerplate



# Semantics

---

- Annotations may be linked to a knowledge base, e.g. DBpedia (<http://dbpedia.org>)
- Annotation refers to an instance
  - [http://dbpedia.org/resource/Harriet\\_Harman](http://dbpedia.org/resource/Harriet_Harman)
- KB knows that this instance belongs to the class of politicians (and people, ...)
  - <http://dbpedia.org/ontology/Politician>
- SPARQL query language can retrieve instances that match constraints

# Semantics

- In news demo, Person, Location and Organization have `class` and `inst` features

```
{Person inst="http://dbpedia.org/resource/Harriet_Harman"}
```

Documents 1 to 20 of 81:

[UK childcare needs to be more affordable - CentreForum](#) (cached)

quality' MP **Harriet Harman** was the architect

[Birth weight among social mobility checks - Nick Clegg](#) (cached)

's deputy leader **Harriet Harman** said Mr Clegg



# Semantic search

---

- Can use SPARQL to query the KB at search time
- E.g. to find *all* politicians

```
{Person sparql="
SELECT DISTINCT ?inst WHERE {
?inst a :Politician }"}
```





# Custom UI

---

- Not the sort of query you want to construct by hand...
- Mimir provides an XML-over-HTTP query API to allow programmatic querying
- Can build custom UIs that hide the query language from users
  - <http://demos.gate.ac.uk/pin>
  - Try out the query builder, look at the underlying query

# Building an index - you need:

---

- Some annotated GATE documents
- A description of which annotations and features you want to index
  - “Index template”
  - Only the features you specify in the template will be available for searching
- A running instance of the Mimir webapp
  - You can download and build your own (requires Grails)
- A way to push the documents to the server



# Pushing documents

---

- GATE PR
  - “Mímir indexing PR” available as part of the Mímir source distribution
- GCP – the “GATE Cloud Parallelizer”
  - Tool to deploy a saved GATE application multi-threaded on your own machine
  - Includes various “output handlers” to save annotations to disk, or push them into Mímir
  - <http://gate.ac.uk/gcp>
- or let us do it for you...



# GATE Cloud

---

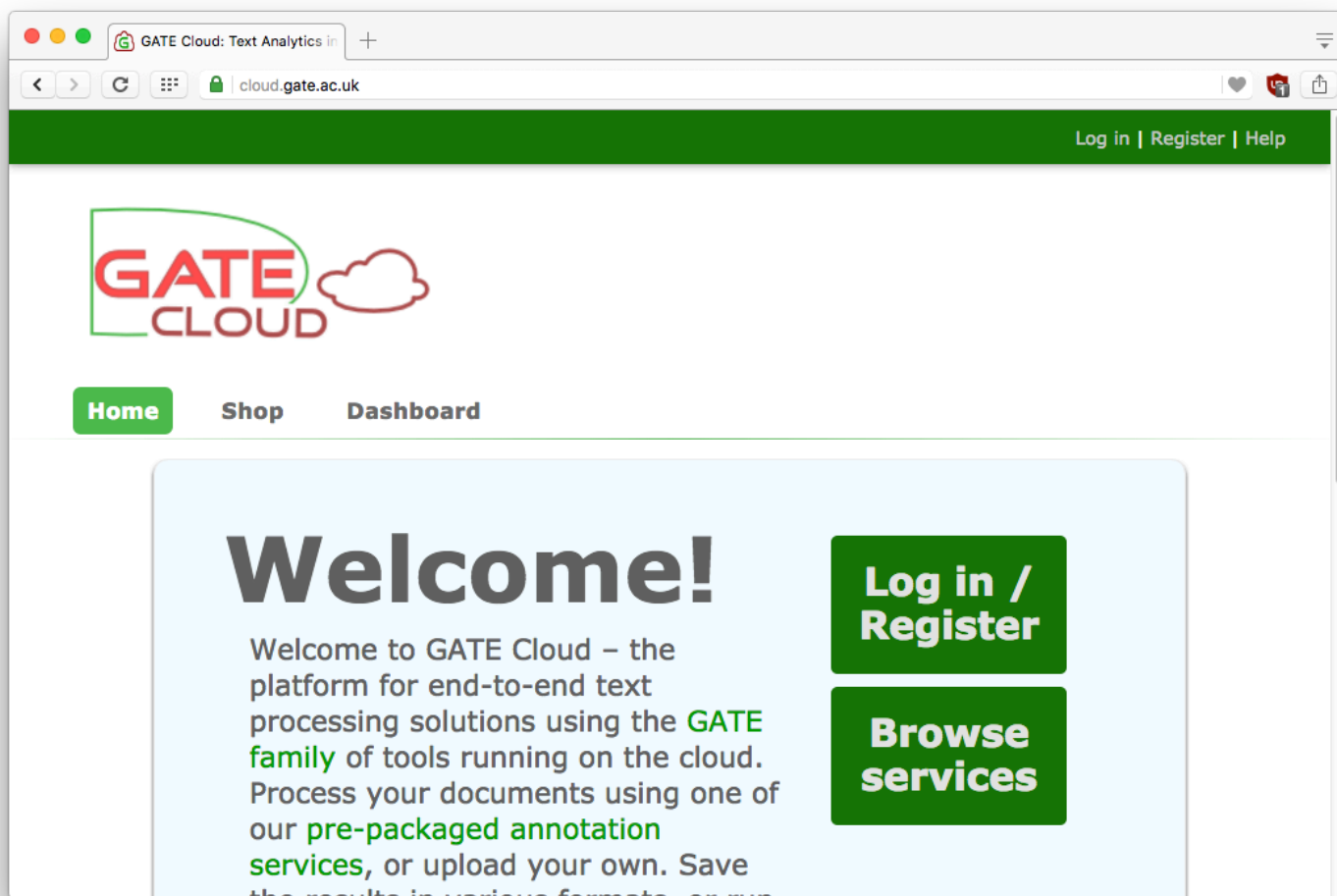
- A cloud based service from the GATE team
- Usual cloud benefits:
  - Pay-as-you-go, no upfront hardware costs
  - No sysadmin work
  - Web-based management tools
  - Always latest version, maintained by us
- Not-so-usual cloud benefits
  - Based on open-source software
  - Bring your own pipeline





# GATE Cloud

- <https://cloud.gate.ac.uk>





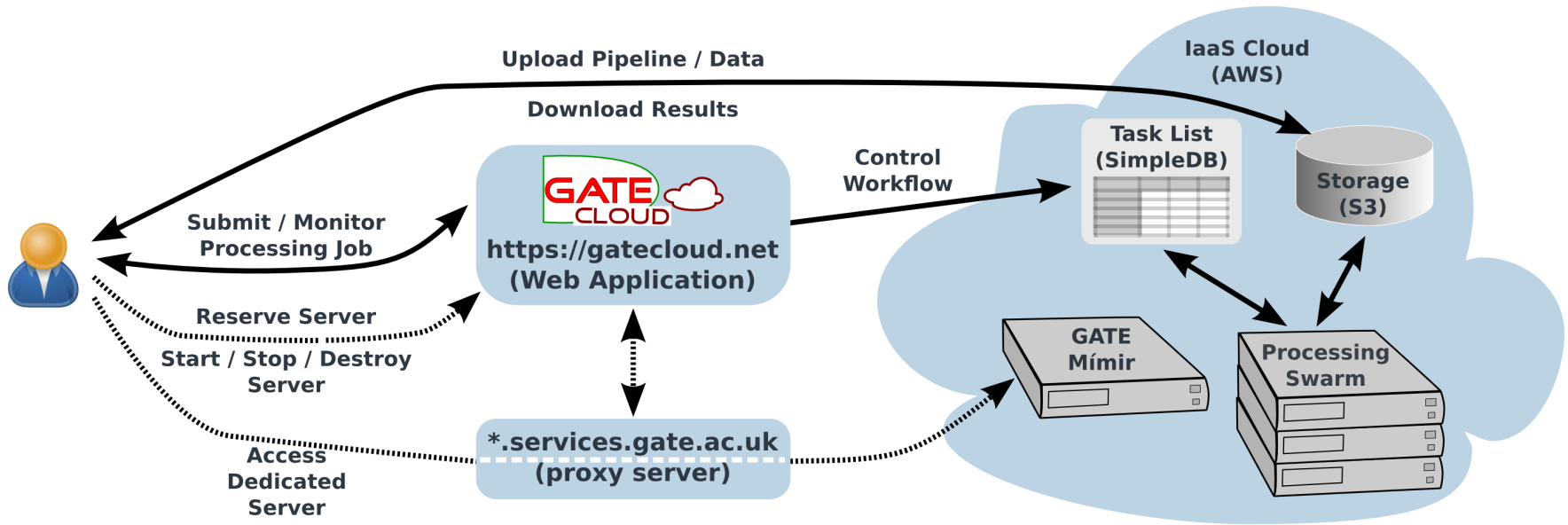
# Features

---

- Pre-packaged pipelines
  - REST API for processing single documents
- On demand document processing (a.k.a. *Annotation Jobs*)
  - Parallel processing on Amazon EC2
  - Using one of our pipelines or your own
  - On-line job definition tool
  - Many output formats, including Mimir
- On demand servers, including Mimir
- Top up your account with vouchers from the University online shop



# Architecture





# Dedicated servers

---


- Rent a dedicated Mimir server for your private use
- Start and stop it as required
- Pay only for the hours it is running
- Data (i.e. indexes) persistent across reboots
- Backup and restore facility available



# The shop

cloud.gate.ac.uk/shopfront


Log in | Register | Help



Home **Shop** Dashboard

Show only items tagged: **Chunker (1)** **Custom (1)** **Dutch (1)** **English (10)** **French (1)** **German (2)** **Measurements (2)**  
**Morphology (1)** **Named Entity (9)** **OpenNLP (3)** **Opinion Mining (1)** **Part-of-Speech (1)** **Spanish (1)** **Summarization (2)**  
**Twitter (4)** **Welsh (1)**


**ANNIE Named Entity Recognizer**



The ANNIE named entity recognition service. Identifies names of *persons, locations, organizations*, as well as *money amounts, time and date expressions*.

**1,200 requests / day**  
**GBP0.80 / CPU hour**


**Custom annotation job**



Upload your own processing pipeline

**Not available**  
**GBP0.80 / CPU hour**


**Mimir Server (5.2)**



Multiparadigm indexing server - index the results of your annotation jobs for rapid retrieval.


**GBP0.50 / CPU hour**

**TwitIE Named Entity Recognizer for Tweets**




Named entity recognition service for Twitter data.

**TwitIE Tweet Tokenizer**



Tokenization service for Twitter data.

**TwitIE Language Identification for Tweets**



Language identification service for Twitter data.



# Reserving a server

---

- The usual e-commerce experience
  - Sign up for an account
  - Buy a top-up voucher
  - Find the server you want in the shop
  - Press “reserve this machine” and follow the instructions
- Server appears in your *dashboard*
- Behind the scenes, creates a persistent data *volume* for your data



# Dashboard

cloud.gate.ac.uk/yourAccount/dashboard

## Ian Roberts: Your Dashboard

### Annotation Jobs

You have no annotation jobs, please visit [the shop](#) to choose the pipeline you want to run.

### Data Bundles

Bundle ID	Name	Created At	Price per month
Upload your own data			

### Cloud Machines

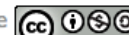
Recently purchased machines may take a few minutes to appear in this list.

[Filter view...](#)

Reservation ID	Machine Type	State	Order Number	Order Date
R-000007	Mimir Server 5.2	inactive	<a href="#">O-000014</a>	June 7, 2016

[About](#)

[Contact](#)





# Reservation control panel

cloud.gate.ac.uk/yourAccount/machineReservationDetails/7

Ian Roberts's account | Log out | Help

The logo for GATE CLOUD, featuring the word "GATE" in red, "CLOUD" in red below it, and a red cloud icon to the right.

Home **Shop** Dashboard

## Machine Reservation R-000007

ID	R-000007	<a href="#">Destroy Reservation</a>
Machine type	Mimir Server (5.2)	
Hourly price	GBP0.50	
State	inactive	<a href="#">Start Instance</a>
Instance ready	no	

### Backups

Slot 1	<empty>	<a href="#">Create new backup</a>
Slot 2	<empty>	<a href="#">Create new backup</a>

## Reservation Details:





# Controlling the server

---

- Start and stop instance
  - Startup/shutdown takes a few minutes – system will email you when server is ready
  - You pay the hourly price whenever the instance is running
- Backup and restore
  - Save the state of your data volume so you can roll back later
- Destroy reservation
  - If you no longer need the server, destroy it to discard the data volume and all backups
  - *This cannot be undone*



# Document processing


---

- Free-of-charge REST API for our standard pipelines
- POST your document, get annotations back
- Quota controlled and rate limited
  - Ask us if you need more quota
- Documentation on GATE Cloud site
- Web form on the shop page to test each pipeline using this API

# “Test this pipeline”

**ANNIE Named Entity Recognizer**

**1,200 requests / day**  
**GBP0.80 / CPU hour**

 **ANNIE** is a named entity recognition pipeline that identifies basic entity types, such as *Person*, *Location*, *Organization*, *Money* amounts, *Time* and *Date* expressions.

It is the prototypical information extraction pipeline distributed with the GATE framework and forms the base of many more complex GATE-based IE applications.

[Annotation details](#)

**Test this pipeline**

Type the content to annotate:

Or select a text file:  No file chosen

Output type:

Document format:

[Customize annotations](#)

[Test Pipeline](#)

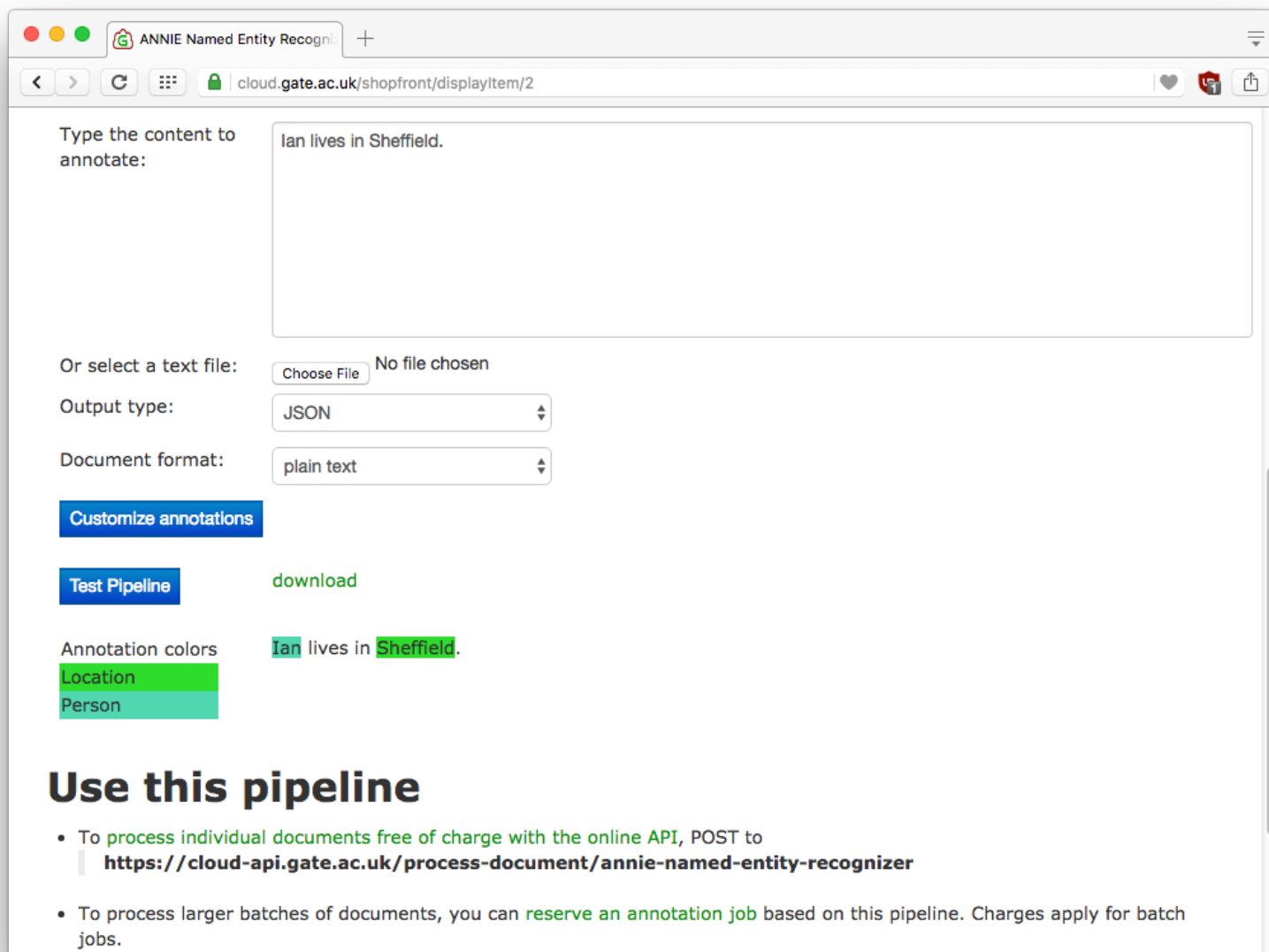
**GATE**

- Applications
  - ANNIE
- Language Resources
  - Processing Resources
    - ANNIE OrthoMatcher
    - ANNIE NE Transducer

Date  
 FirstPerson  
 JobTitle  
 Location  
 Lookup  
 Money  
 Organization  
 Person  
 Sentence  
 SpellingRules  
 Tokens

Deutsche Telekom yesterday delayed the demerger of its mobile wireless business, which owns UK network O2zone, until next year owing to the volatility of the stock markets.  
 Deutsche Telekom was among a number of European telecom companies looking to demerge or float wireless operations this year. It is expected to demerge its wireless business in the future.  
 However, Deutsche Telekom, still partly state-owned, warned yesterday that investor appetite for hi-tech shares is still too low. Analysts consider T-Mobile, which has 60m subscribers across Europe, to be one of Deutsche Telekom's most valuable assets and welcomed the decision to delay.

# Pipeline results

A screenshot of a web browser showing the ANNIE Named Entity Recognition interface. The browser address bar shows 'cloud.gate.ac.uk/shopfront/display/item/2'. The main content area has a text input field with 'Ian lives in Sheffield.' and a 'Choose File' button. Below are dropdown menus for 'Output type' (JSON) and 'Document format' (plain text). There are buttons for 'Customize annotations', 'Test Pipeline', and 'download'. A legend shows 'Location' in green and 'Person' in blue. The text 'Ian lives in Sheffield.' is shown with 'Ian' highlighted in blue and 'Sheffield' in green. A section titled 'Use this pipeline' contains two bullet points about API usage.

Type the content to annotate:

Or select a text file:  No file chosen

Output type:

Document format:

[Customize annotations](#)

[Test Pipeline](#) [download](#)

Annotation colors

Location	Ian lives in Sheffield.
Person	

## Use this pipeline

- To process individual documents free of charge with the online API, POST to <https://cloud-api.gate.ac.uk/process-document/annie-named-entity-recognizer>
- To process larger batches of documents, you can [reserve an annotation job](#) based on this pipeline. Charges apply for batch jobs.



# Calling the API

---

- You can call the API directly via HTTPS
- Generate an API key from your account page
  - The key consists of an identifier (username) and secret (password), used for basic authentication
- POST to endpoint, with appropriate parameters
  - <https://cloud.gate.ac.uk/info/help/online-api.html>



# Examples

---

- Curl

```
curl -H "Accept: application/json" -H  
"Content-Type: text/plain" --data-binary  
"Ian lives in Sheffield" -u  
"apiKey:password" https://cloud-  
api.gate.ac.uk/process-document/annie-  
named-entity-recognizer
```

- Java client library available from GitHub & our Maven repo
  - <https://github.com/GateNLP/cloud-client>
- Plugin to call the API from GATE Developer
- Other languages have similar tools
  - e.g. Python “requests” module



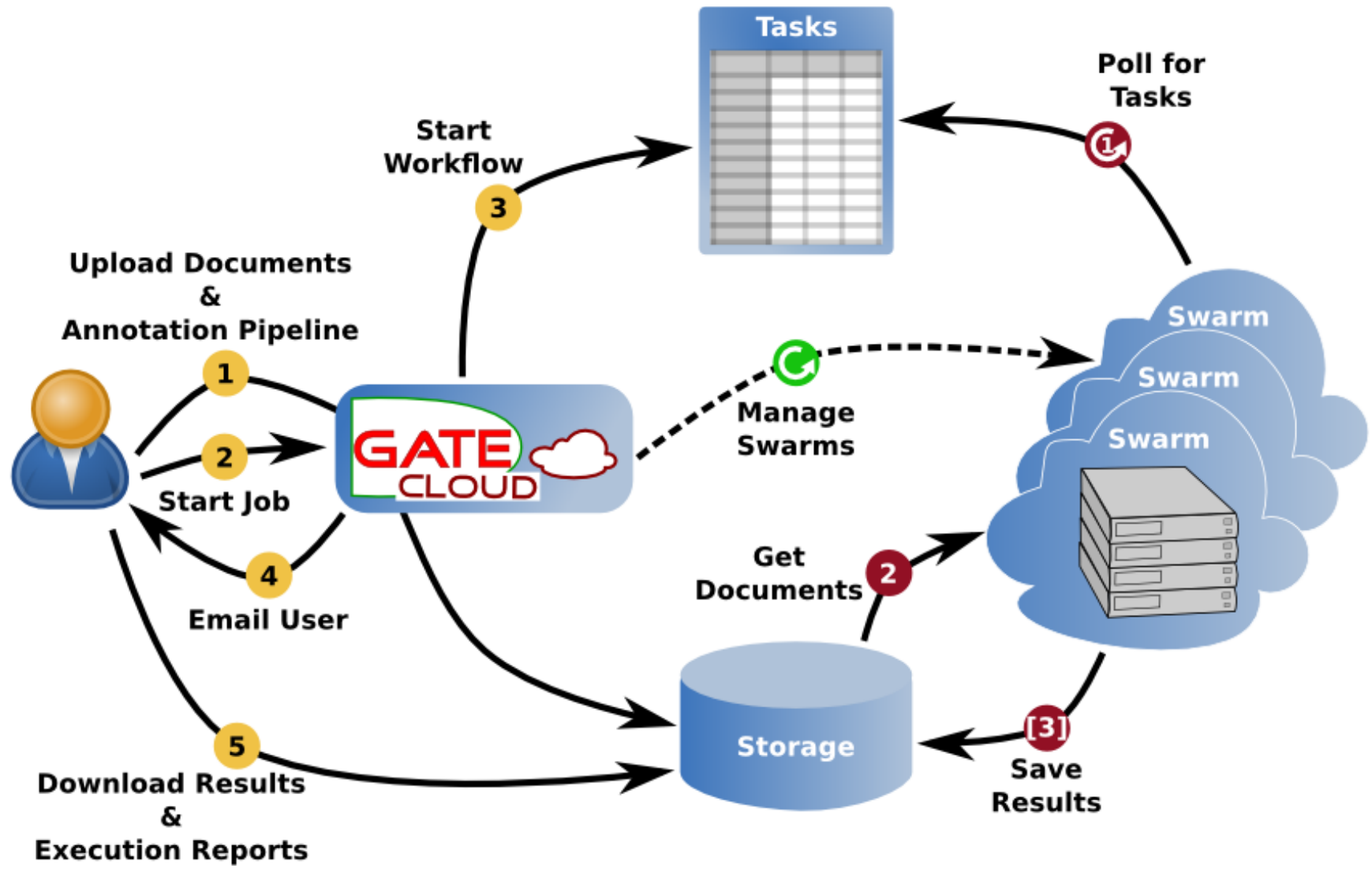
# Annotation jobs

---

- For larger amounts of data
- Parallel and distributed annotation of documents with a GATE application
  - Choose from the standard pipelines
  - Or upload your own pipeline
    - “Export for GATECloud.net”
- Upload your own documents
  - zip, tar, arc/warc archives, Twitter JSON
  - Or collect data directly from Twitter API (later)
- Output annotations in various formats, or send documents directly to Mimir



# Job lifecycle







# Execution environment

---

- Amazon EC2
- Ubuntu LTS, 64-bit
- Oracle Java 8
- ~2GB/thread RAM on average
- GCP 2.8.1, based on GATE Embedded 8.4.1

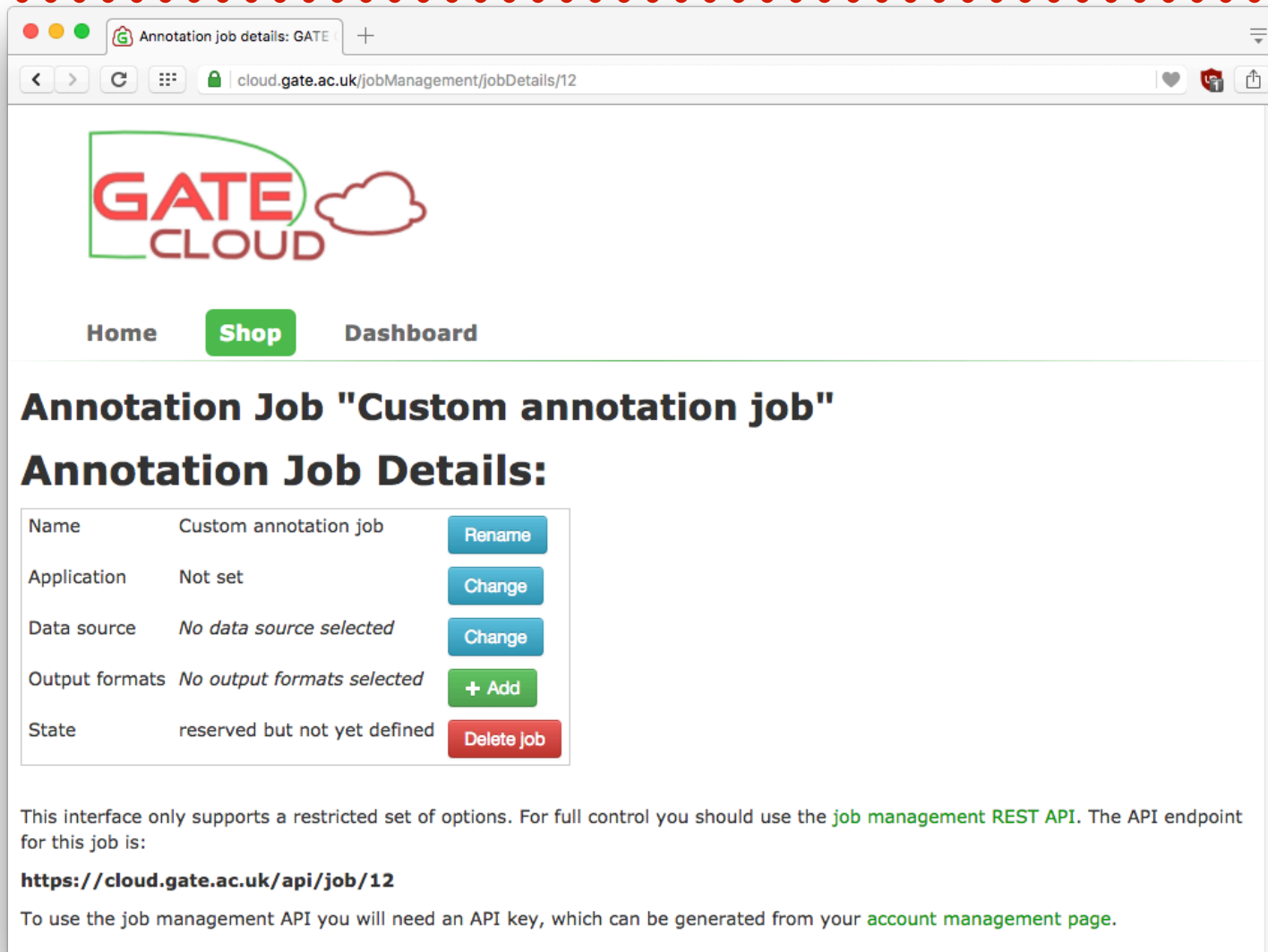


# Reserving a job

---

- Use the “reserve an annotation job” link from the pipeline page
- Or select “custom annotation job” to provide your own pipeline
  - This is what we will do here, since we need the morphological analyser as well as the basic ANNIE
- Once reserved, jobs appear in your dashboard

# Managing a job



The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The page features the GATE Cloud logo and navigation links for Home, Shop, and Dashboard. The main content area displays the title "Annotation Job 'Custom annotation job'" and "Annotation Job Details:". Below this is a table of job properties with associated action buttons.

Name	Custom annotation job	<a href="#">Rename</a>
Application	Not set	<a href="#">Change</a>
Data source	<i>No data source selected</i>	<a href="#">Change</a>
Output formats	<i>No output formats selected</i>	<a href="#">+ Add</a>
State	reserved but not yet defined	<a href="#">Delete job</a>

This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is:

**<https://cloud.gate.ac.uk/api/job/12>**

To use the job management API you will need an API key, which can be generated from your [account management page](#).



# Managing a job - application

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The page title is "Annotation job details: GATE". A modal dialog box titled "Upload application" is open, showing the text "Application zip file: Choose File annie-with-morph.zip". Below the text are "Cancel" and "Upload" buttons. The background page is dimmed and shows a table with job details:

Name	Custom annotation
Application	annie-with-morph.zip <span>Change</span>
Data source	No data source selected <span>Change</span>
Output formats	No output formats selected <span>+ Add</span>
State	reserved but not yet defined <span>Delete job</span>

At the bottom of the page, there is a note: "This interface only supports a restricted set of options. For full control you should use the job management REST API. The API endpoint..."



# Managing a job - input

---

- Annotation job input and output files are modelled as “data bundles”
  - Persistent sets of data
  - You pay per GB per month for storage
  - Delete the bundle when you no longer require it
- Bundle can be created by uploading files
- Output of a job becomes another bundle
  - So you can feed the output of one job to the input of another



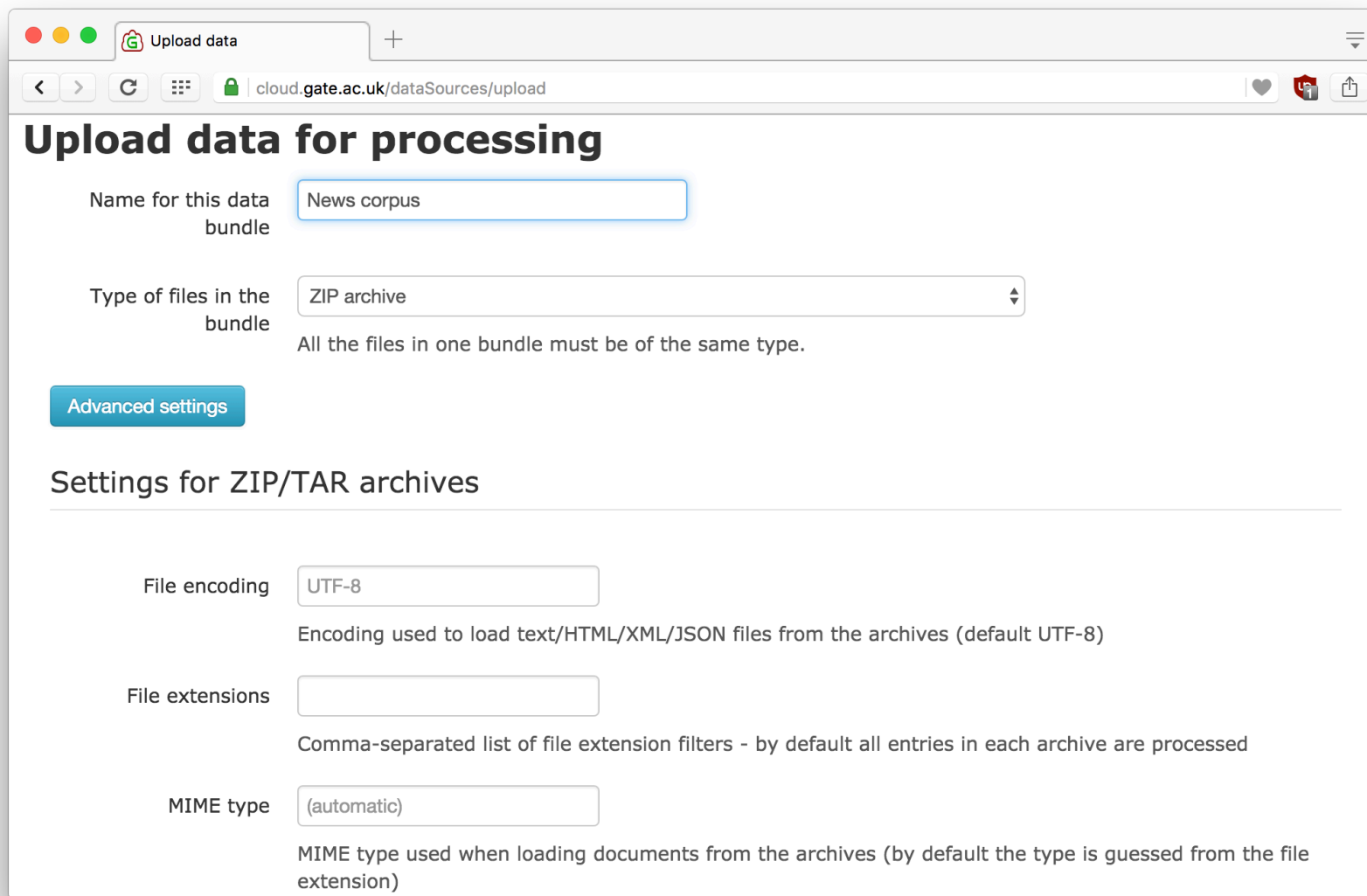
# Managing a job - input

A screenshot of a web browser window showing the GATE job management interface. The browser tab is titled "Annotation job details: GATE" and the address bar shows "cloud.gate.ac.uk/jobManagement/jobDetails/12". The main content area displays "Annotation Job 'Custom annotation job'" with a table of job details. A modal dialog box titled "Select data source" is overlaid on the page, containing the text: "No suitable data sources found. You can also [upload your own data](#), or use the API if you have an existing data set on Amazon S3". A "Cancel" button is located at the bottom right of the dialog. The background table shows the following details:

Name	Custom an...
Application	annie-with-
Data source	No data so...
Output formats	No output formats selected
State	reserved but not yet defined

At the bottom of the page, there is a link: <https://cloud.gate.ac.uk/dataSources/upload> and a paragraph of text: "options. For full control you should use the job management REST API. The API endpoint".

# Managing a job - input

A screenshot of a web browser window showing the 'Upload data for processing' page. The browser's address bar shows 'cloud.gate.ac.uk/dataSources/upload'. The page has a title 'Upload data for processing' and contains several form fields and a button. The 'Name for this data bundle' field contains 'News corpus'. The 'Type of files in the bundle' dropdown is set to 'ZIP archive'. Below this is a note: 'All the files in one bundle must be of the same type.' There is a blue button labeled 'Advanced settings'. Below that is a section titled 'Settings for ZIP/TAR archives' with three form fields: 'File encoding' set to 'UTF-8', 'File extensions' (empty), and 'MIME type' set to '(automatic)'. Each field has a descriptive text below it.

Upload data

cloud.gate.ac.uk/dataSources/upload

## Upload data for processing

Name for this data bundle

Type of files in the bundle

All the files in one bundle must be of the same type.

[Advanced settings](#)

### Settings for ZIP/TAR archives

File encoding   
Encoding used to load text/HTML/XML/JSON files from the archives (default UTF-8)

File extensions   
Comma-separated list of file extension filters - by default all entries in each archive are processed

MIME type   
MIME type used when loading documents from the archives (by default the type is guessed from the file extension)



# Managing a job - input

The screenshot shows a web browser window with the following content:

- Browser Tab:** Data bundle D-000011
- Address Bar:** cloud.gate.ac.uk/dataManagement/dataBundleDetails/11
- Section Header:** Data bundle D-000011
- Details Panel:**

ID	D-000011	
Name	News corpus	<a href="#">Rename</a>
Date created	June 7, 2016 9:47:47 PM BST	
Monthly price	GBP0.03	
Type	ZIP archive	
Input mime type	text/html	
Actions	<a href="#">Delete this bundle</a>	

This will also delete the underlying files
- Section Header:** Upload data
- Text:** You can add more data to this bundle - when you have finished adding files, click the "Finished" button
- Text:** Upload a ZIP archive:
- Form:**
- Form:**
- Text:** 0 files added so far.





# Managing a job - input

A screenshot of a web browser window. The address bar shows the URL 'cloud.gate.ac.uk/dataManagement/dataBundleDetails/11'. The page title is 'Data bundle D-000011'. A green notification bar at the top of the content area says 'Data bundle D-000011 closed successfully'. Below this is a table of bundle details with a 'Rename' button. At the bottom of the details box is a 'Delete this bundle' button. Below the details box, it says 'This will also delete the underlying files' and 'This bundle contains 1 data files.' A table below shows one file: 'news-corpus-large.zip'. The footer contains 'About' and 'Contact' links.

ID	D-000011	
Name	News corpus	<a href="#">Rename</a>
Date created	June 7, 2016 9:47:47 PM BST	
Total size	18.09 MiB	
Monthly price	GBP0.03	
Type	ZIP archive	
Input mime type	text/html	
Actions	<a href="#">Delete this bundle</a>	

This will also delete the underlying files

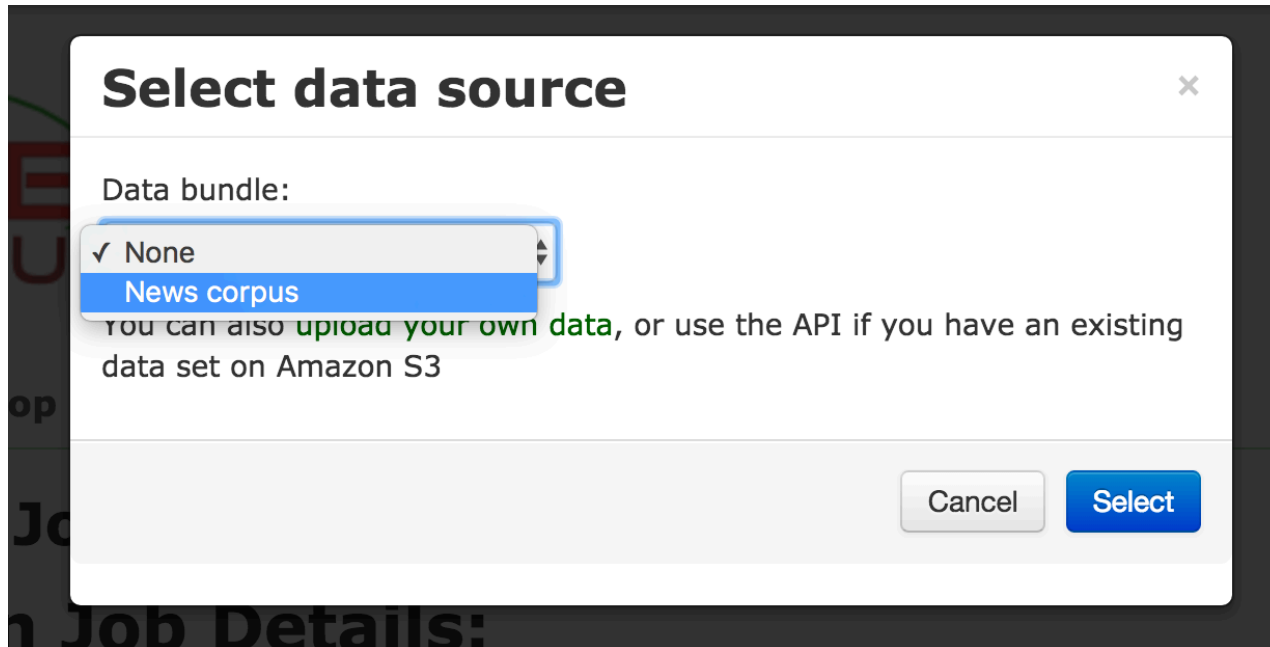
This bundle contains 1 data files.

Filename
news-corpus-large.zip

[About](#) [Contact](#)

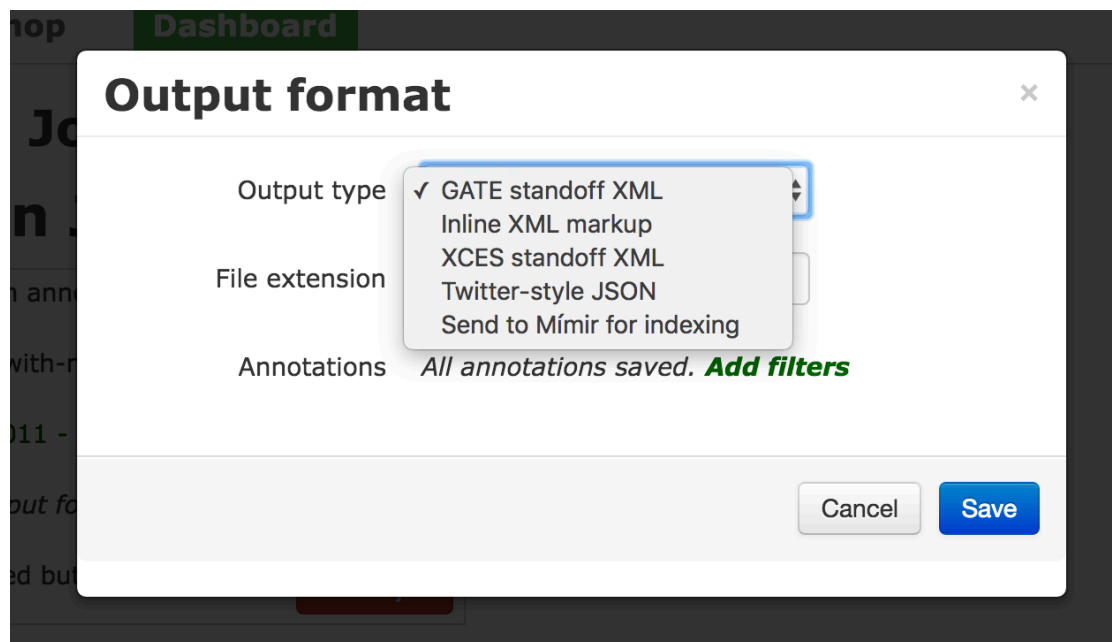
# Managing a job - input

- Back at the job editor



# Managing a job - output

- Various output formats supported
  - Save annotated documents as GATE XML
  - Twitter-style JSON
  - Index in Mimir



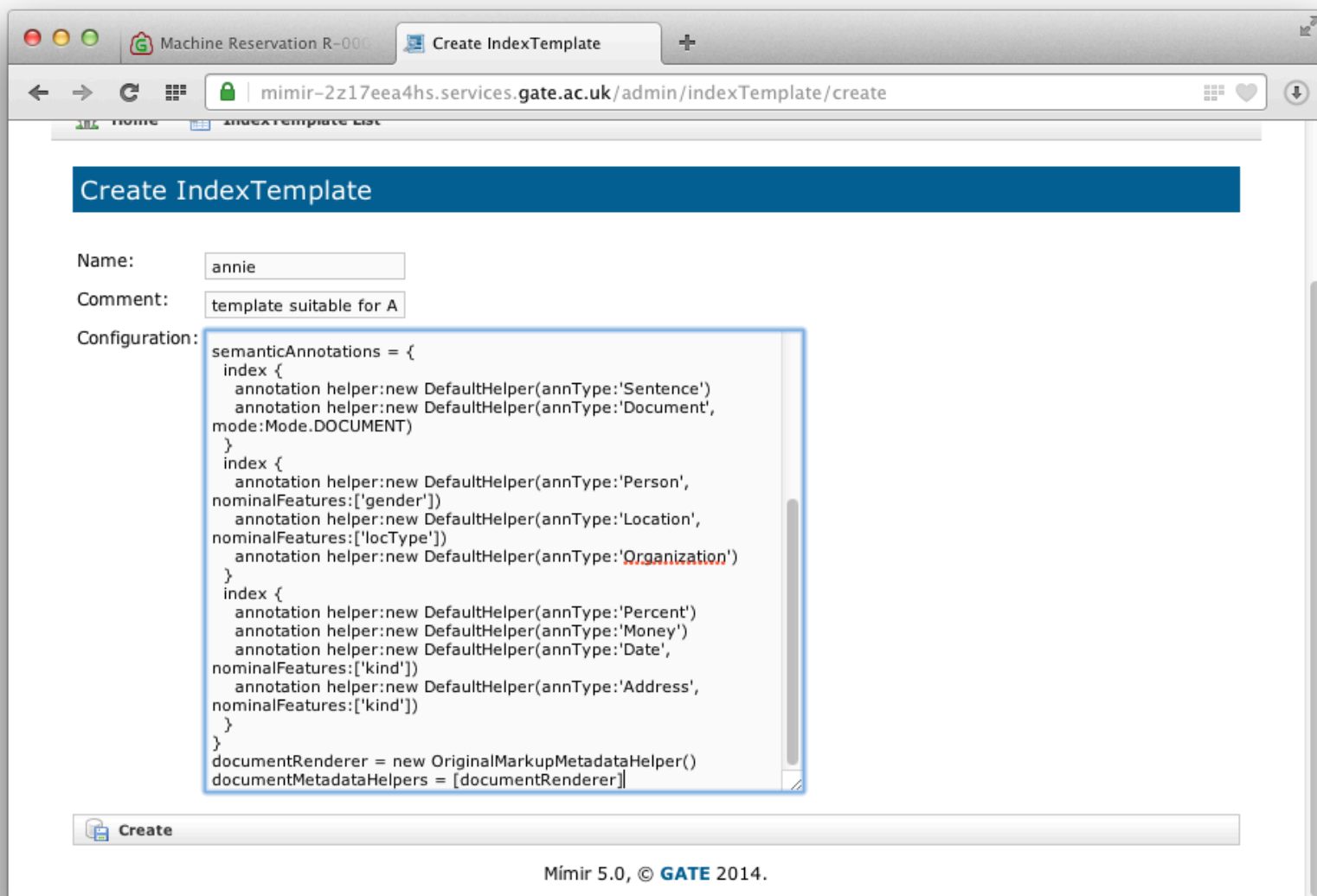


# Outputting to Mimir

---

- Start up the Mimir server we reserved earlier
- Create an appropriate index template
- Create an index

# Create a template



The screenshot shows a web browser window with the address bar displaying `mimir-2z17eea4hs.services.gate.ac.uk/admin/indexTemplate/create`. The page title is "Create IndexTemplate".

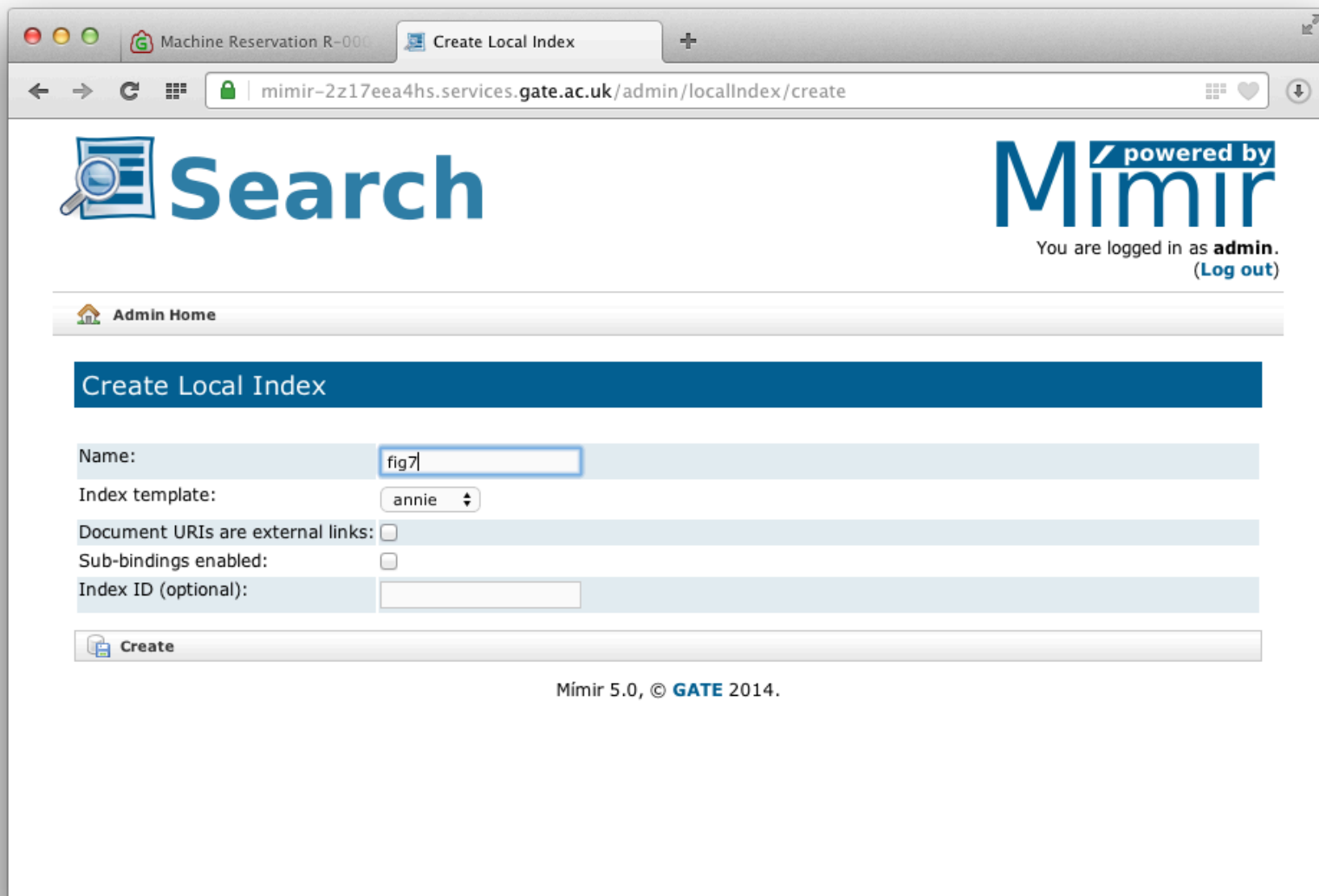
The form contains the following fields:

- Name:
- Comment:
- Configuration: 

```
semanticAnnotations = {  
  index {  
    annotation helper:new DefaultHelper(annType:'Sentence')  
    annotation helper:new DefaultHelper(annType:'Document',  
mode:Mode.DOCUMENT)  
  }  
  index {  
    annotation helper:new DefaultHelper(annType:'Person',  
nominalFeatures:['gender'])  
    annotation helper:new DefaultHelper(annType:'Location',  
nominalFeatures:['locType'])  
    annotation helper:new DefaultHelper(annType:'Organization')  
  }  
  index {  
    annotation helper:new DefaultHelper(annType:'Percent')  
    annotation helper:new DefaultHelper(annType:'Money')  
    annotation helper:new DefaultHelper(annType:'Date',  
nominalFeatures:['kind'])  
    annotation helper:new DefaultHelper(annType:'Address',  
nominalFeatures:['kind'])  
  }  
}  
documentRenderer = new OriginalMarkupMetadataHelper()  
documentMetadataHelpers = [documentRenderer]
```


At the bottom of the form is a "Create" button. The footer of the page reads "Mimir 5.0, © GATE 2014."

# Create an index

A screenshot of a web browser window showing the Mimir admin interface. The browser's address bar shows the URL 'mimir-2z17eea4hs.services.gate.ac.uk/admin/localIndex/create'. The page features a 'Search' logo on the left and 'powered by Mimir' on the right, with a user login status 'admin' and a '(Log out)' link. Below the navigation bar is a blue header for 'Create Local Index'. The main form contains several fields: 'Name' with the value 'fig7', 'Index template' with a dropdown menu showing 'annie', 'Document URIs are external links' with an unchecked checkbox, 'Sub-bindings enabled' with an unchecked checkbox, and 'Index ID (optional)' with an empty text box. At the bottom of the form is a 'Create' button. The footer of the page reads 'Mimir 5.0, © GATE 2014.'


Machine Reservation R-000 Create Local Index

mimir-2z17eea4hs.services.gate.ac.uk/admin/localIndex/create

 **Search**

powered by  
**Mimir**

You are logged in as **admin**.  
([Log out](#))

 [Admin Home](#)

## Create Local Index


Name:

Index template:

Document URIs are external links:

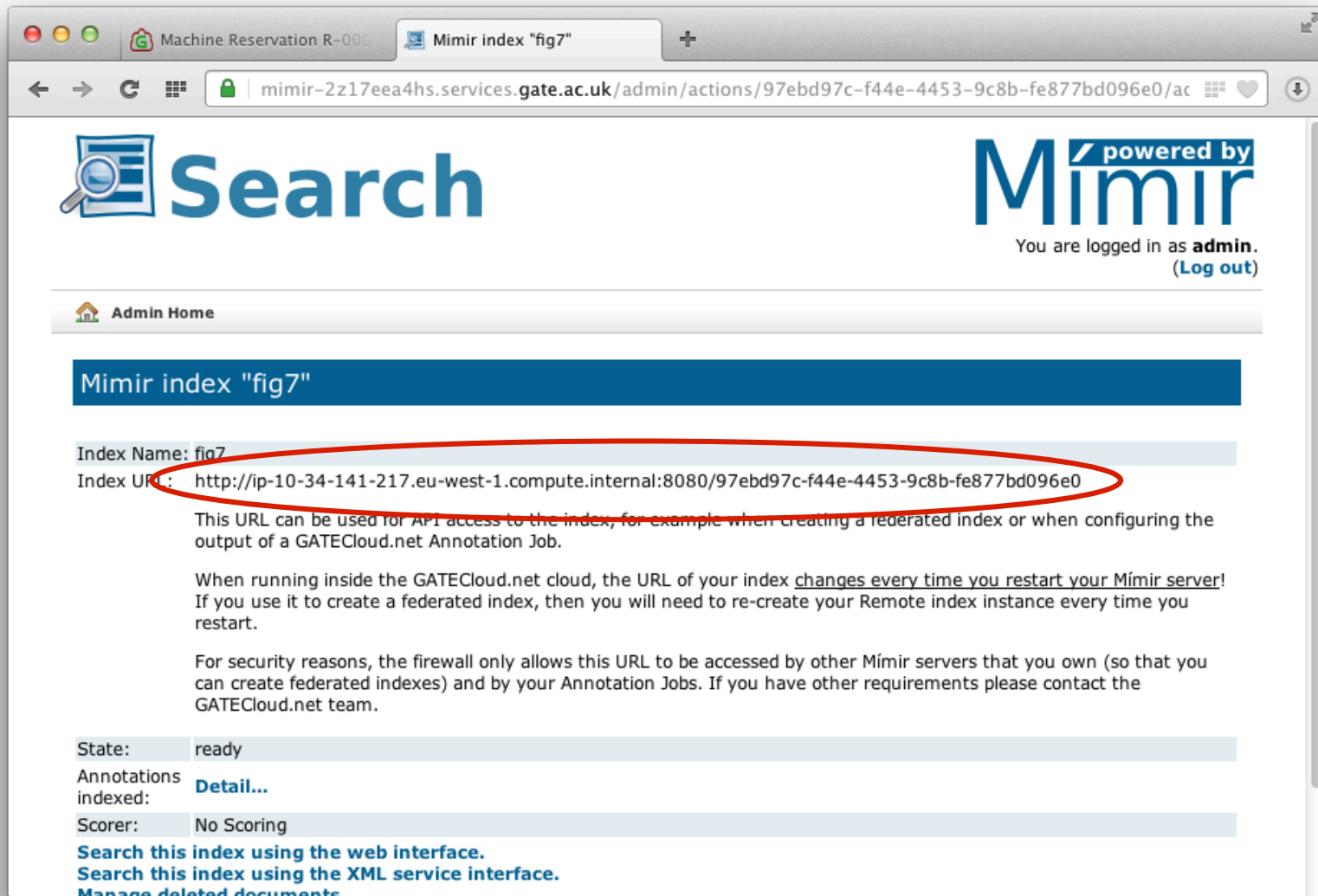
Sub-bindings enabled:

Index ID (optional):

 **Create**

Mimir 5.0, © GATE 2014.

# Index details – note URL



The screenshot shows a web browser window with the URL `mimir-2z17eea4hs.services.gate.ac.uk/admin/actions/97ebd97c-f44e-4453-9c8b-fe877bd096e0/ac`. The page features a search bar with a magnifying glass icon and the text "Search". To the right, it says "powered by Mimir" and "You are logged in as admin. (Log out)". Below this is a navigation bar with "Admin Home". The main content area is titled "Mimir index 'fig7'".

Index Name: fig7

Index URL: <http://ip-10-34-141-217.eu-west-1.compute.internal:8080/97ebd97c-f44e-4453-9c8b-fe877bd096e0>

This URL can be used for API access to the index, for example when creating a federated index or when configuring the output of a GATECloud.net Annotation Job.

When running inside the GATECloud.net cloud, the URL of your index changes every time you restart your Mimir server! If you use it to create a federated index, then you will need to re-create your Remote index instance every time you restart.

For security reasons, the firewall only allows this URL to be accessed by other Mimir servers that you own (so that you can create federated indexes) and by your Annotation Jobs. If you have other requirements please contact the GATECloud.net team.

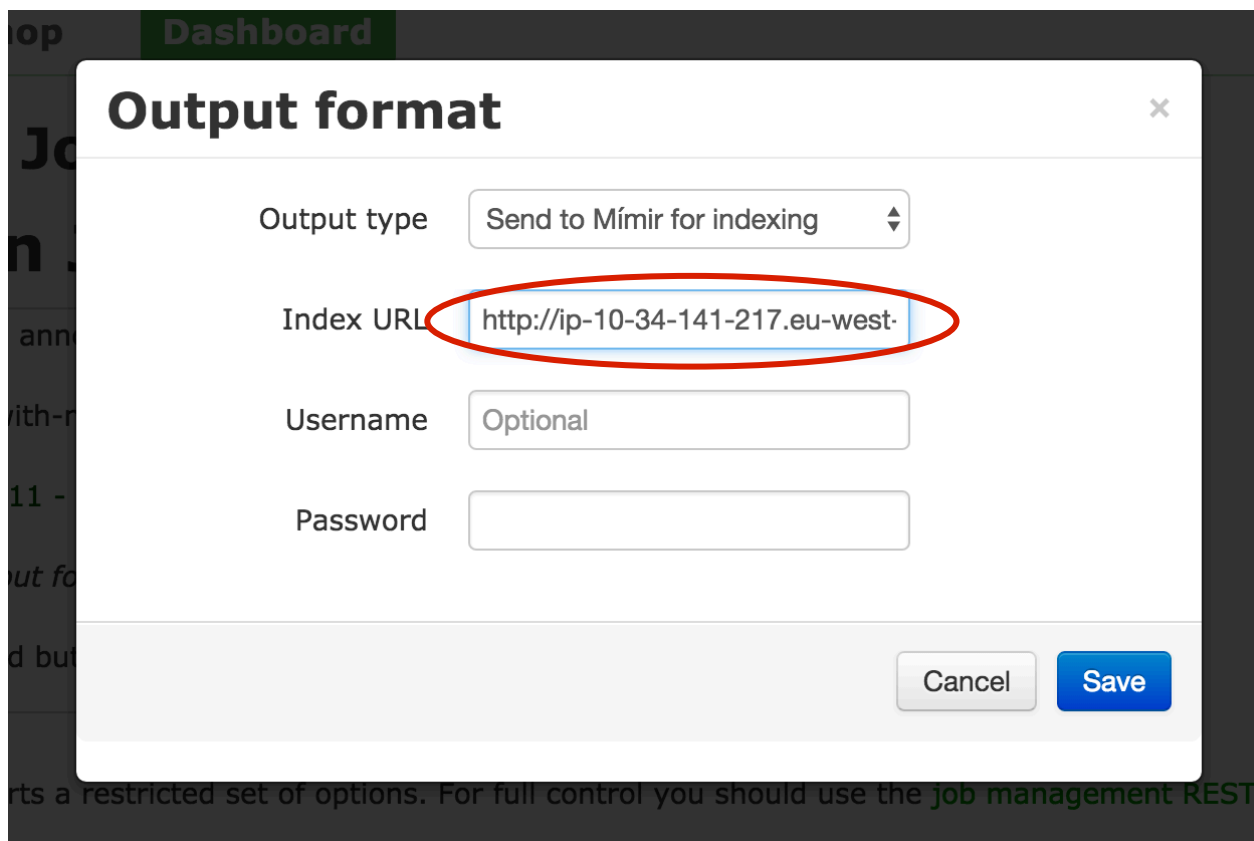
State: ready

Annotations indexed: [Detail...](#)

Scorer: No Scoring

[Search this index using the web interface.](#)  
[Search this index using the XML service interface.](#)  
[Manage deleted documents.](#)

# Managing a job - output

A screenshot of a web application interface showing a dialog box titled "Output format". The dialog box has a close button (X) in the top right corner. It contains four input fields: "Output type" with a dropdown menu showing "Send to Mimir for indexing"; "Index URL" with a text input field containing "http://ip-10-34-141-217.eu-west-" and a red oval highlighting the text; "Username" with a text input field containing "Optional"; and "Password" with an empty text input field. At the bottom right of the dialog box are two buttons: "Cancel" and "Save". The background of the application shows a "Dashboard" tab and some text fragments like "op", "Jo", "n.", "ann", "with-r", "11 -", "ut fo", "d bu", and "rts a restricted set of options. For full control you should use the job management REST".

**Output format** ×

Output type

Index URL

Username

Password





# Start the job

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The page title is "Annotation job details: GATE". The navigation menu includes "Home", "Shop", and "Dashboard". The main heading is "Annotation Job 'Custom annotation job'" followed by "Annotation Job Details:". A table lists job details with corresponding action buttons:

Name	Custom annotation job	Rename
Application	annie-with-morph.zip	Change
Data source	D-000011 - News corpus	Change
Output formats	Send to Mimir for indexing	Edit Remove
		+ Add another
State	ready to be started	Start job Delete job

The "Start job" button is circled in red. Below the table, a note states: "This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is: **<https://cloud.gate.ac.uk/api/job/12>** To use the job management API you will need an API key, which can be generated from your [account management page](#)."



# Job running

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The navigation menu includes **Home**, **Shop**, and **Dashboard**. The main heading is **Annotation Job "Custom annotation job"**. A green notification bar states "Job successfully started". Below this, the **Annotation Job Details:** section contains a table with the following information:

Name	Custom annotation job	<a href="#">Rename</a>
Application	annie-with-morph.zip	
Data source	D-000011 - News corpus	
Output formats	Send to Mimir for indexing	
State	active - processing	<a href="#">Interrupt job</a>

Below the table is a progress bar showing 33.33333333299999995% completion.

This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is:

**<https://cloud.gate.ac.uk/api/job/12>**

To use the job management API you will need an API key, which can be generated from your [account management page](#).



# When job completes...

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The page has a navigation bar with 'Home', 'Shop', and 'Dashboard' (highlighted in green). The main heading is 'Annotation Job "Custom annotation job"'. A green notification bar states 'Job successfully started'. Below this is the 'Annotation Job Details:' section, which contains a table of job information and several action buttons.

Name	Custom annotation job	<a href="#">Rename</a>
Application	annie-with-morph.zip	
Data source	D-000011 - News corpus	
Output formats	Send to Mimir for indexing	
State	completed	<a href="#">Reset job</a> <a href="#">Delete job</a>
Final report	<a href="#">Download</a>	
Debug logs	<a href="#">Download</a>	

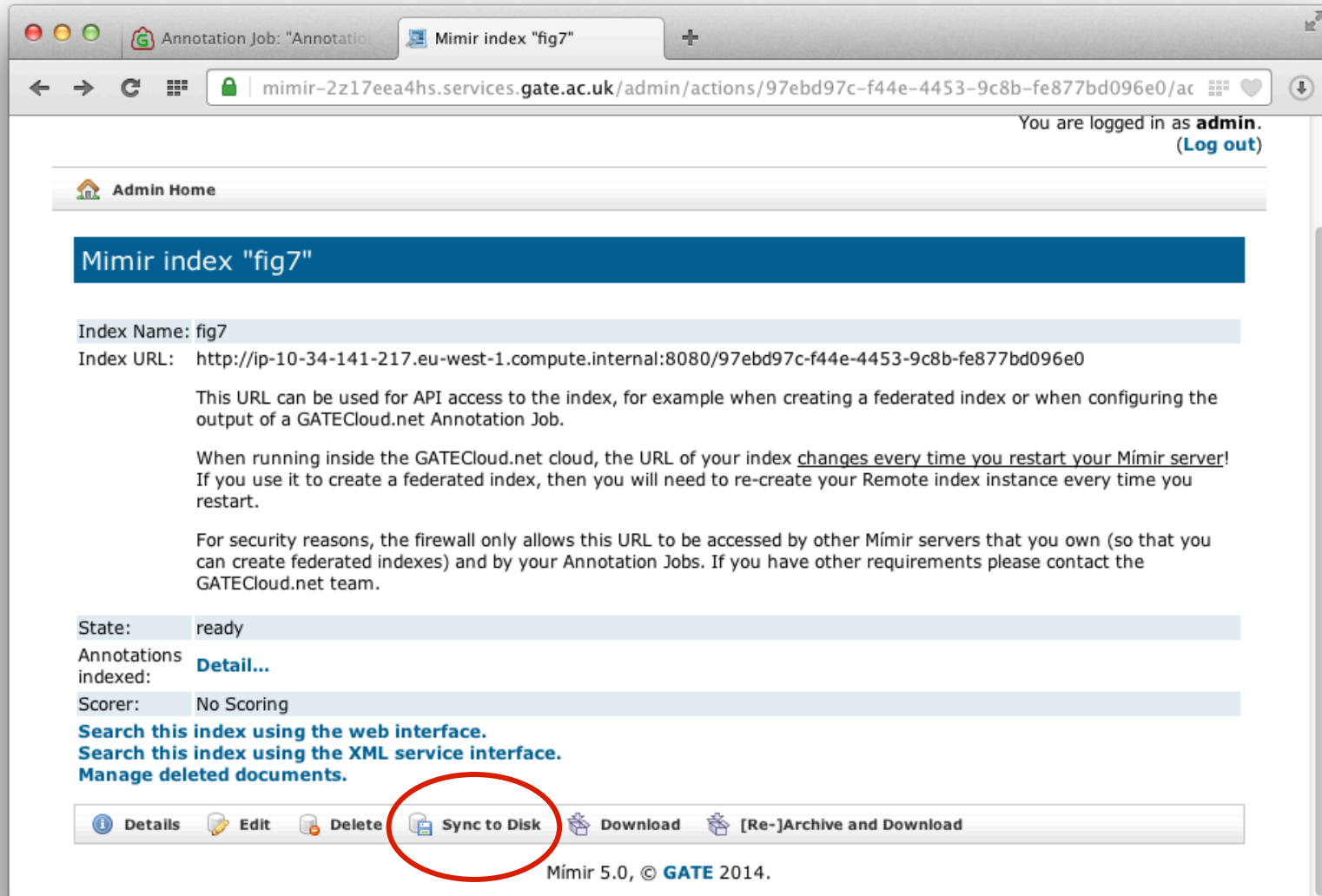
This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is:

**<https://cloud.gate.ac.uk/api/job/12>**

To use the job management API you will need an API key, which can be generated from your [account management page](#).

Waiting for cloud.gate.ac.uk...

# “Sync” the index



The screenshot shows a web browser window with the URL `mimir-2z17eea4hs.services.gate.ac.uk/admin/actions/97ebd97c-f44e-4453-9c8b-fe877bd096e0/ac`. The user is logged in as **admin.** (Log out). The page title is **Mimir index "fig7"**. The index name is `fig7` and the index URL is `http://ip-10-34-141-217.eu-west-1.compute.internal:8080/97ebd97c-f44e-4453-9c8b-fe877bd096e0`. The index state is **ready**. The interface includes a toolbar with buttons for **Details**, **Edit**, **Delete**, **Sync to Disk** (circled in red), **Download**, and **[Re-]Archive and Download**. The footer indicates **Mimir 5.0, © GATE 2014.**

Annotation Job: "Annotation" Mimir index "fig7"

You are logged in as **admin.** (Log out)

**Admin Home**

## Mimir index "fig7"

Index Name: fig7  
Index URL: `http://ip-10-34-141-217.eu-west-1.compute.internal:8080/97ebd97c-f44e-4453-9c8b-fe877bd096e0`

This URL can be used for API access to the index, for example when creating a federated index or when configuring the output of a GATECloud.net Annotation Job.

When running inside the GATECloud.net cloud, the URL of your index changes every time you restart your Mimir server! If you use it to create a federated index, then you will need to re-create your Remote index instance every time you restart.

For security reasons, the firewall only allows this URL to be accessed by other Mimir servers that you own (so that you can create federated indexes) and by your Annotation Jobs. If you have other requirements please contact the GATECloud.net team.

State: ready  
Annotations indexed: [Detail...](#)  
Scorer: No Scoring

[Search this index using the web interface.](#)  
[Search this index using the XML service interface.](#)  
[Manage deleted documents.](#)

[Details](#) [Edit](#) [Delete](#) [Sync to Disk](#) [Download](#) [\[Re-\]Archive and Download](#)

Mimir 5.0, © GATE 2014.



# “Sync” the index

---

- Mimir accumulates documents in RAM
- Documents saved to disk after (by default) one hour, or when memory threshold reached
- Documents become searchable once saved to disk
- “Sync” button forces an immediate save, if you know no more documents due
  - Can continue to send more documents, but only sync-ed ones available for search



# Search your new index

A screenshot of a web browser window showing the Mimir search interface. The browser's address bar contains the URL: `mimir-2z17eea4hs.services.gate.ac.uk/97ebd97c-f44e-4453-9c8b-fe877bd096e0/search/index#qu`. The page features a search bar with a magnifying glass icon and the word "Search" in large blue letters. To the right, it says "powered by Mimir" and "You are logged in as admin. (Log out)". A blue banner indicates "Searching Index 'fig7'". Below this, a text input field contains the query `{Person gender = "female"}`, and a "Search" button is positioned to its right. A second blue banner shows "Documents 1 to 20 of 760:". The search results list includes a Guardian article from 2009: `www.guardian.co.uk/business/2009/dec/20/shoppers-defy-big-freeze?INTCMP=ILCNETTXT3487` and a book listing: "39 5. Sarah Raven's Wild Flowers Wild Flowers by Sarah Raven £29." Another Guardian article from 2008 is partially visible at the bottom.



# Try it!

- Sign up for an account on <https://cloud.gate.ac.uk>
- Use your voucher code
- Reserve a Mimir 5.3 server
- Start it up, log in
- Create a new index template using the contents of `index-template.groovy`
- Create a new local index using this template
- Visit index admin page and note the URL



# Try it!

- Reserve a “custom annotation job”
- Application zip file is annie-with-morph.zip
- Create a data bundle for the input, and upload news-corpus-large.zip
  - Mime type: text/html, Encoding: UTF-8





# Try it!

- Set one output to MIMIR, using the Index URL you noted above
  - Make sure to not include any spaces in the index URL
- Run the job, and when finished sync the index



# Try it!

- Try some searches on your new index
  - E.g. stock price movements  
{Organization} (up | down) ({Money} | {Percent})
- When finished, make sure you stop the Mimir server, destroy the reservation, and delete the data bundle you created



# Collecting data from Twitter

---

- Tool to collect tweets in real time from Twitter's streaming API
- Collect based on keyword/hashtag search, follow specific users, or geo-locations
- Tweets stored in your dashboard as a daily data bundle
  - You can process these with annotation jobs, just like your own uploaded data
- Detailed documentation and video tutorials on GATE Cloud site



# Questions?

---

- More info
  - <https://cloud.gate.ac.uk>
  - <https://gate.ac.uk/mimir>